トピックモデルによる話題知識を考慮したテンプレート穴埋め型発話生成

著者	久保田 豊久
内容記述	筑波大学修士(情報学)学位論文 ・ 平成29年3月24日授与(37753号)
発行年	2017
URL	http://hdl.handle.net/2241/00150786

トピックモデルによる話題知識を考慮したテンプレート穴埋め型発話生成

筑波大学

図書館情報メディア研究科

2017年3月

久保田 豊久

目 次

図 目 次

表 目 次

第1章　序論

近年，ロボット技術や携帯端末の発達とともに，対話システムの実装が人々の身近に現れたことで，その高度化に注目が集まっている．自然言語対話によって様々な形式の情報を伝えることができるが，その中でも特定の目的を持たない対話である雑談は，人間同士の対話全体の約 5 割を占めることが報告されている [1]．人間が自然に対話システムを利用する上で，システムが雑談に対応する機能をもつことは，システムに対する信頼感の向上や，話者の潜在的な情報要求の発見において重要な役割を果たすと考えられる．このことから，特定の話題に限定されないオープンドメインな雑談対話に対応できるシステムへの期待は，関連の産業において高まっているといえる．

オープンドメインな対話システムとして，これまでには，web 上のテキストから関連度の高い文章を選択する手法 [2][3] や，人手で作成したルールに基づいた返答を行う対話システムの構築手法 [4] が提案されている．ルールベースの手法は，設定したルールで想定された範囲内の形式や話題の対話を行う上では自然な対話を行うことができるが，対応可能な話題を増やすとルールが難解になっていくという問題を抱えている．一方，web テキストの選択に基づく対話手法は多くの話題に対応できるが，web テキストの集合は有限であるのに対して，自然言語対話は状況ごとに無限に多様であるため，状況と web テキストとの間の「ずれ」が本質的に解消できないという問題を持つ．

本研究では，発話の形式と話題を分離し，それらを組み合わせて状況に合わせた発話を生成する方式の手法を検討する．安藤ら [5] は，生成方式の対話システムの実現を目指し，テンプレートを用いた発話生成手法を提案している．これは，過去の会話履歴データにある発話文の一部を空欄に置き換えることで作成した発話テンプレートを用いて，発話生成時に空欄に適切な語を当てはめることで，文構造を保ちつつ，様々な話題に適応可能な手法である．しかし，この提案手法では，テンプレートを選択する基準がなくランダムに選ばれることや，空欄に当てはめる単語の選択方法において前後の単語や文脈との連接関係を考慮していないといった問題がある．本論文では，テンプレートを用いた発話文生成手法において，Support Vector Machine (SVM) を用いた発話タイプ推定に基づいたテンプレート選択と，N-gram モデルによる当てはめ単語の連接関係の考慮，トピックモデルを用いた話題を考慮した単語選択により，より尤もらしいオープンドメイン発話文の生成を行う手法を提案する．実験により，提案手法と従来手法の比較を行い，有用性を検証する．

第2章　関連研究

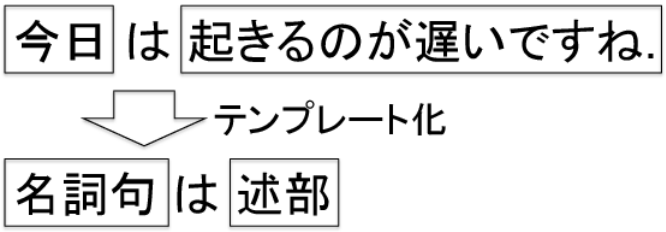

図 2.1: テンプレート化

安藤ら [5] は，特定のチャットルームでの会話履歴を用例文集合として用い，テンプレートに基づいた発言モデルによるテキストコミュニケーションを行う対話システムを構築する手法を提案している．発言モデルは，用例文集合から構築した空欄を含む発話テンプレートの集合から 1 つを選び，別途用例文集合から抽出した語彙用例集から選んだ語彙列を空欄に当てはめることで発話文生成を行う．テンプレートの空欄は,「名詞句」と「述部」の 2 種類とする．テンプレートの構築ではまず用例文に対して形態素解析を行い，名詞が連続している部分単語列を「名詞句」，最初に出現した動詞から文末までの部分単語列を「述部」と置き換えて，テンプレートとする．テンプレートの例を図 2.1 に示す．語彙用例集は用例文集合から抽出した連接辞書であり，それぞれの単語について，次の位置に出現したことのある単語のリストが登録される．発話文の生成ではランダムに選択されたテンプレートについて，それぞれの空欄に置換語句を当てはめる．置換語句の生成は以下の手順で行う．

- 空欄が名詞句ならば名詞を，述部ならば動詞を，語彙集合からランダムに選択し，1 番目の単語とする．
- $n = 2, 3, 4, ...$ について，以下を繰り返す．
  - 語彙用例集から，$n-1$ 番目の単語の次に出現しうる単語のリストを参照し，その中からランダムに単語を 1 つ選ぶ．選んだ単語を n 番目の単語とする．
  - 空欄が名詞句ならば名詞，述部ならば句読点が終了語彙となる．選択された n 番目の単語が終了語彙ならば，ループを抜けて終了する．

この手法は名詞句と述部のランダムな置き換えに基づいて生成がされており，意味の通らない文が生成されることが多いが，安藤らは改善を施す余地が十分にあると述べている．手法の

利点として，文構造を保ちつつ豊富な種類の文生成が行える点にあるが，置換語句選択する際に，テンプレートとの連接関係の考慮や話題の考慮が出来ていないという問題を持つ．提案手法では N-gram モデルとトピックモデルを用いることで，連接関係と話題を考慮した置換語句の選択を行い，文生成を行うことでこれらの課題を解決することを検討する．

第3章　生成手法

先行研究では，名詞句と述部という空欄を設定することでテンプレート生成を行っていたが，本研究では単純化のために名詞のみを空欄に設定した．文の生成に際しては，テンプレート選択方法として Support Vector Machine (SVM)[6][7] による発話タイプ推定を用い，N-gram モデル，トピックモデル，両モデルを組み合わせた提案手法によって文生成を行う．

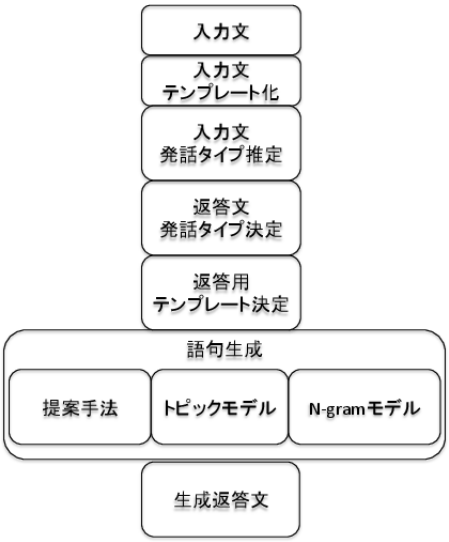

図 3.1: 生成手順概略

3.1　提案手法の概要

テンプレート穴埋め発話生成は，テンプレートを選択するステップと，空欄を穴埋めする置換語句を選択するステップに分けられる．処理の概略を図 3.1 に示す．

テンプレート選択のステップでは，テンプレート同士の隣接関係を考慮した方法と，発話タイプを考慮した方法を適用する．コーパス全体をテンプレート化する際に，テンプレートの隣接関係を記録しておき，入力文テンプレートと全く同一のテンプレートがコーパス中に存在する場合，コーパス中でその次の返答に用いられているテンプレートを選択する．入力

文のテンプレートがコーパス中に存在しない場合，入力文の発話タイプの推定を行い，その返答の発話タイプにふさわしいと考えられる発話タイプをもつテンプレートをランダムに選択する．発話タイプの推定に基づくテンプレート選択については 3.2 節で述べる．

テンプレート選択後の置換語句選択ステップにおいては，トピックモデルのみを用いた手法，N-gram のみを用いた手法，およびトピックモデルと N-gram の両方を考慮した提案手法のそれぞれで置換する語を選択する．トピックモデルの学習については 3.3 節で，それぞれの置換語句の選択手法については 3.4 節で述べる．

3.2 発話タイプの推定

テンプレート選択のステップでは，大まかな発話の意図が決定すると考えられることから，入力文の発話タイプの推定に基づいてテンプレートの選択を行う．発話タイプは SWBD-DAMSL タグ [8] を独自に統廃合し，表 3.1 に示すように，Statement-opinion (sv)，Wh-Question (qw)，Other answers (no)，Other-forward-function (fo)，Uninterpretable (%) の５つを用いる．タグの統合に際しては Yes/No 回答が可能な場合に付与される質問タグ，Yes/No 回答タグ等を質問，回答タグとして統合を行っている．同意や相づちを示すタグについても利用率が小さいために挨拶等を示す fo タグに統合，もしくは廃止を行っている．

本研究では Twitter から抽出した会話データに含まれる 800 会話について，人手で上記の 5 種類の発話タイプのいずれかを付与するアノテーションを行った．アノテーションされたデータは Support Vector Machine (SVM)[6][7] の訓練データとして用いられる．訓練データはテンプレートに変換され，テンプレートに含まれる全単語の単語頻度ベクトルを特徴ベクトルとして SVM の学習が行われる．発話タイプ推定は学習を行った SVM を用い，コーパス全体から構築したテンプレートそれぞれについて推定を行う．発話生成を行う際には，入力文の発話タイプを推定し，あらかじめ表 3.2 のように定めた発話タイプ遷移表に従って，入力文の発話タイプの返答としてふさわしいと考えられる発話タイプをランダムに一つ選択する．さらに選択した発話タイプをもつテンプレートを一つランダムに選択し，返答文のテンプレートとする．表 3.2 の遷移表はアノテーション済み会話データの発話タグ遷移を確認しつつ，複数人の話し合いを通じて人手で作成した．

SVM の実装には，scikit-learn の LinearSVC を用いた．

表 3.1: SWBD-DAMSL 統合タグ

SWBD-DAMSL タグ	内容
Statement-opinion (sv)	意見
Wh-Question (qw)	質問
Other answers (no)	回答
Other-forward-function (fo)	挨拶等
Uninterpretable (%)	解釈不能

表 3.2: 発話タイプ遷移表

入力発話タイプ	出力発話タイプ
sv	sv,qw,fo
qw	no,fo
no	sv,qw,fo
fo	sv,qw,fo
%	%

3.3 トピックモデルの学習

本研究では，トピックモデルとして Latent Dirichlet Allocation (LDA) を用いる [9]．LDA は，Bag-of-words で表現された文書の集合 $X = \{X_1, ..., X_D\}$， $X_d = \{x_{d1}, ..., x_{dN_d}\}$ に関する確率的生成モデルである．文書 d に含まれる個々の単語 x_{di} は，トピックと呼ばれる潜在変数 z_{di} に依存して生成される．各文書は個別のトピック分布 θ_d を潜在的にもつ．また，トピック k に対応する単語の確率分布を ϕ_k とする．LDA では，これらの変数の同時分布について，以下の条件付き独立性を仮定する．

$$
\begin{aligned}
p(X, Z, \theta, \phi | \alpha, \beta) &= \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(x_{di} | \phi_{z_{di}}) p(z_{di} | \theta_d) \\
&\times \prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{k=1}^{K} p(\phi_k | \beta_k)
\end{aligned} \tag{3.1}
$$

ここで，K はトピックの数である．$p(\theta_d|\alpha)$ および $p(\phi_k|\beta_k)$ は，それぞれパラメータ α，β_k のディリクレ分布である．$\beta = \{\beta_1, ...\beta_K\}$ とする．

LDA の学習とは，所与の文書集合 X に対して，尤もらしい Z, θ, ϕ の値を推定することである．本研究では，LDA の学習アルゴリズムとして Mimno ら [10] が提案した確率的変分ベイズ法を用いる．この手法は確率的最適化に基づいており，反復最適化の 1 回のイテレーションごとに文書集合 X から B 件の文書をランダムサンプリングして得られたミニバッチを用いて学習を行う．ここで，B はバッチサイズと呼び，本研究では $B = 4000$ とした．この手法により，大規模な文書集合の学習を行う場合でもスケーラブルに学習を行うことが可能である．

このアルゴリズムにより，対話コーパスを 1 対話 1 文書とみなした文書集合を用いて LDA の学習を行う．学習により，各トピックの単語生成確率分布 ϕ が得られる．また，これを用いることで，未知の入力文章に対して，トピック分布 $\bar{\theta}$ を推論することができる．

3.4 置換語句の選択

本節では，選択されたテンプレートにおける置換語句を選択する手法として，トピックモデルのみを用いた手法，マルコフ連鎖のみを用いた手法およびトピックモデルとマルコフ連鎖を組み合わせた手法についてそれぞれ述べる．トピックモデルを用いる手法の場合には，まず，ユーザの入力文を単語に分割し，上記で学習した LDA を用いて入力文のトピック分布 $\bar{\theta}$ を求める．

トピックモデルのみを用いた手法では，前後の連接関係を無視して，置換語句 x の確率分布を以下のように展開する．

$$p(x|\bar{\theta}) = \sum_{k=1}^{K} p(x|z=k)p(z=k|\bar{\theta})$$

ここで，$p(z=k|\bar{\theta})$ は $\bar{\theta}$ をパラメータとする離散分布であるから，$p(z=k|\bar{\theta}) = \bar{\theta}_k$ である．また，$p(x|z=k)$ は，トピック k が単語 x を生成する確率であるから，コーパスで学習されたパラメータを用いて $p(x|z=k) = \phi_k$ である．

マルコフ連鎖のみを用いた手法では，コーパスから学習した 2 次のマルコフ連鎖を用いて，置換語句 x の前 2 単語 x_{-2}, x_{-1} および後 2 単語 x_{+1}, x_{+2} を用いて以下のように x のスコアを求める．

$$\begin{aligned} &p(x|x_{-2}, x_{-1}, x_{+1}, x_{+2}) \\ &\propto p(x|x_{-2}, x_{-1})p(x_{+1}|x_{-1}, x)p(x_{+2}|x, x_{+1}) \end{aligned}$$

ここで，$p(x|x_{-2}, x_{-1}), p(x_{+1}|x_{-1}, x), p(x_{+2}|x, x_{+1})$ はそれぞれコーパスから最尤推定した 2 次のマルコフモデルで推定される遷移確率である．

提案手法である，トピックモデルとマルコフ連鎖を組み合わせた手法では，置換語句 x 自体はトピックモデルから生成され，続く後 2 単語 x_{+1}, x_{+2} が 2 次のマルコフモデルによって生成されると仮定する．置換語句 x のスコアは以下のように求める．

$$\begin{aligned} &p(x|x_{-1}, x_{+1}, x_{+2}) \\ &\propto \sum_{k=1}^{K} p(x|z=k)p(z=k|\bar{\theta})p(x_{+1}|x_{-1}, x)p(x_{+2}|x, x_{+1}) \end{aligned}$$

それぞれの手法で，全ての語彙についてスコアを求め，その中で最もスコアの大きい語彙を置換語句として選択する．

第4章　実験

4.1　実験方法

N-gram のみを用いた文生成，トピックモデルのみを用いた文生成，提案手法による文生成をそれぞれ行い，実験結果に対してアンケート調査を実施することで，評価を行う．実験では 2013 年投稿において，リプライが連鎖しているようなツイートを会話として抽出した対話コーパス (約 150MB) から返答文生成を行う．また Twitter 日本語対話文 800 対話に発話タイプアノテーションを施した，発話タイプ推定のための学習コーパスを用いて入力文及び出力文の発話タイプ推定を行う．入力文，入力文発話タイプ推定，出力文発話タイプ推定，発話テンプレート選択は共通とし，生成部分のみを比較する．比較評価はアンケート調査によって行い，アンケートはクラウドソーシングサービス「Lancers」を通じて行う．

4.2　評価方法

アンケートでは表 4.1 のように入力文と選択肢が与えられ，最も相応しいと思われる回答を選択する．返答文 A は提案手法によって生成された文，返答文 B はトピックモデルのみを用いて生成された文，返答文 C は N-gram モデルのみを用いて生成された文である．返答文 D は回答者がどれも適切な文でないと判断した場合に選択する．入力文として 10 種類の文章を作成し，それぞれの入力文について 8 回ずつ，各手法で返答文を生成し，計 80 種類のアンケートを作成した．各アンケートにつき 30 人の回答者に回答してもらい，合わせて 24000 回答を得た．

表 4.1: アンケート例

入力文	今日も暇だな
返答文 A	今日遊ぼうか
返答文 B	こと遊ぼうか
返答文 C	こんど遊ぼうか
返答文 D	どれも適切でない

4.3 実験結果

表 4.2 のアンケート回答結果を見ると，全回答 24000 回答の内，N-gram モデルのみを用いた生成結果を支持した回答が 2024 回答，トピックモデルのみを用いた生成結果を支持した回答が 555 回答，提案手法を用いた生成結果を支持した回答が 1173，どれも適切でないとした回答が 20248 回答となった．まず,「どれも適切でない」の支持は全体の約 84%を占めたことから，今回の実験方法による生成には多くの改善余地が残されていると考えられる.「どれも適切でない」を選択肢から除いた場合には，N-gram モデルのみを用いた生成結果が全体の約 54%を占める支持を獲得し，提案手法を用いた生成結果は約 31%，トピックモデルのみを用いた生成結果は約 15%の支持を得た．回答者が入力文に対して適切な生成文であると回答した設問の内，およそ半数は N-gram のみを用いた生成手法を支持し，手法を併用している提案手法を合わせた場合には，全体の約 85%が連接関係を考慮した手法を支持したという結果となった．一方で，話題を考慮すべく導入されたトピックモデルによる生成は提案手法を合わせた場合でも全体の約 45%の支持を得るに留まるという結果となった.

表 4.2: アンケート回答結果

手法	回答数（全 24000 回答）	全体から占める割合
提案手法	1173	0.0489
トピックモデル	555	0.0231
N-gram モデル	2024	0.0843
どれも適切でない	20248	0.843

次に個々の生成結果を見ると，表 4.3 の生成例ではダイエットに関する文が入力文として選択され，返答に使用されるテンプレートは「[　　] ですか？」という疑問を返答するテンプレートが選択されている．発話タイプの推定では，入力文は意見を表す sv タグが推定，返答文には質問を表す qw タグが推定されている．トピックモデルのみを用いた場合の生成では,「今日ですか？」という文が生成された．これは入力文の「最近」という語が時間に関する語であり,「今日」という語も時間に関する語であることから，入力文のトピックにおいて時間に関係するトピックの成分が大きくなり,「今日」という語が選択されたと考えられる．生成文単体としては確認を取るような出力となっているが，入力文と見比べると意味は通っていない．N-gram モデルのみを用いた場合の生成文は,「俺ですか？」という文が生成された．N-gram 辞書を作成する際，文頭に第一人称が存在する文が数多く見受けられ，同時にですます調の文もよく見られていた．そのため連接としても登場しやすく,「俺」という第一人称を表す語が選択されたと考えられる．生成文単体を見た場合，確認を取るような文出力となっており，入力は自身がダイエットを始めたことを示す文であることから,「俺ですか？」では意味が通らない回答になっている，提案手法を用いた場合の生成文では,「愚痴ですか？」という文が生成された．これは学習コーパスにおいて「ダイエット」という語の周辺に「愚痴」というがある程度存在していたことから，同じ文脈において登場しやすい単語であり，単語トピック分布が近くなったことで選出されたと考えられる．生成文単体で見た場合には，確

認を取るような生成文であり，ダイエットを始めたことが愚痴であるかのように確認する文の出力のように見て取ることができる．

表 4.4 の生成結果では，旅行に関する文が入力文として選択され，返答に使用されるテンプレートは「[　　] だっけ？」という疑問を返答するテンプレートが選択されている．発話タイプの推定では，入力文は意見を表す sv タグが推定され，返答文には質問を表す qw タグが推定されている．トピックモデルのみを用いた場合の生成文では,「火曜日だっけ？」という文が生成された．生成結果は表 4.3 と同様に「今年」という語が時間に関する語であり,「火曜日」という語も曜日という時間に関係する語であったためにトピックを表す語として選出されたと考えられる，生成された文と入力文を見比べると,「ハワイに行った」という意見に対して，火曜日に行ったのかという質問で返しているように見えるが,「今年の夏」という時期を明記した文が存在しているために，質問として成り立っておらず，返答として意味が通っていない文となっている．N-gram モデルのみを用いた場合の生成では,「何だっけ？」という文が生成された.「何」という語はコーパス全体に登場する頻出語であったために，辞書生成時に特に出現しやすい語となってしまったと考えられる，そのために返答文は入力文に対して，意味の通らない文となってしまっている．提案手法を用いた場合の生成では,「写真だっけ？」という文が生成された.「ハワイに行く」という文面から入力文は旅に関連するトピック成分が大きくなると考えられる．トピックモデルのみを用いた場合には時間に関するトピックが選出されていたが，この生成文では「写真」という旅に関連すると考えられる語が選出されている．これは，トピックが近い語の候補を選出した段階で時間に関するトピック以外にも文の話題に沿ったトピックが選出される可能性を示していると考えられる．

表 4.5 の生成結果では，入力文に月曜日の感想が綴られ，返答に使用されるテンプレートには「[　　] かい？」という疑問を返答するテンプレートが選択されている．発話タイプの推定では，入力文は意見を表す sv タグが推定され，返答文には質問を表す qw タグが推定されている．トピックモデルのみを用いた場合の生成は入力文に含まれる,「月曜日」という語が曜日，時間に関するトピックを含んでいると考えられることから，前例の生成例と同様に「今日」という時間に関する語が選出されてしまっている．さらに生成された文は入力文への応答として，意味の通らない文となっている．N-gram モデルのみを用いた場合の生成では,「つらい」という文面に対して「大丈夫かい？」という心配を表明する文を生成しており，返答としても十分の意味の通った文が生成されている.「大丈夫かい？」という文は相手のこと心配する文脈で使用されることが容易に想定できることから,「つらい」という語の返答として一定以上の頻度でコーパス内に登場していたと考えられる．提案手法を用いた場合の生成では,「朝」という「月曜日」と共通した時間に関する語かつ，コーパス内において同じ文，もしくは近い文で同時に利用されやすいであろう語が選出されている．トピックと連接共に関連している語が選出されているが，返答としては意味の通らない文となってしまっている．

表 4.6 の生成結果では，雑談に関する文が入力文として選択され，返答に使用されるテンプレートは「 [　　] は寝るね、おやすみ！」という意見表明を行うテンプレートが選択されている，発話タイプの推定では，入力文は意見を表す sv タグが推定され，返答文にも意見を表す sv タグが推定されている．トピックモデルのみを用いた場合の生成では，入力文に「寝る

ね，おやすみ」という睡眠や時間に関係する語が出現したために，これまでの生成例と同様に時間に関係する語である,「今」が選出されている．提案手法による生成もまた「今日」という語を選出しており，時間に関するトピック成分が大きかったのではないかと考えられる．N-gram モデルのみを用いた場合の生成では，第一人称を表す「僕」という語が選択されており，連接として登場しやすい語の並びであったと考えられる，どの生成文も文単体としては意味が通っているが，入力文に対しては意味が通らない文となっている．入力文に対して適切なテンプレート文が選択されなかったことが原因と考えられるが，これはテンプレートを作成した時点で文脈がほぼ固定されてしまうという，テンプレート集生成時の問題であるとも考えられる．

表 4.3: 個別生成例 1

入力文	最近、ダイエット始めたんだ
選択テンプレート	[　　] ですか？
提案手法生成文	愚痴ですか？
トピックモデル生成文	今日ですか？
N-gram モデル生成文	俺ですか？

表 4.4: 個別生成例 2

入力文	今年の夏はハワイに行って来たよ
選択テンプレート	[　　] だっけ？
提案手法生成文	写真だっけ？
トピックモデル生成文	火曜日だっけ？
N-gram モデル生成文	何だっけ？

表 4.5: 個別生成例 3

入力文	毎度の如く月曜日がつらい
選択テンプレート	[　　] かい？
提案手法生成文	朝かい？
トピックモデル生成文	今日かい？
N-gram モデル生成文	大丈夫かい？

表 4.6: 個別生成例 4

入力文	雑談できるなんてホントかよ
選択テンプレート	[　　] は寝るね、おやすみ！
提案手法生成文	今日は寝るね、おやすみ！
トピックモデル生成文	今は寝るね、おやすみ！
N-gram モデル生成文	僕は寝るね、おやすみ！

4.4　実験考察

評価アンケートの結果より，語の連接関係を考慮することが尤もらしさに寄与し，大きな支持を得たことが確認された．N-gram モデルのみを用いた場合には前後 2 形態素の連接関係を考慮していたが，提案手法は後方 2 形態素のみの連接考慮となったために，十分に文脈を考慮出来なかったと考えられる．ただし N-gram モデルを使用した場合でも，話題を提供しない「何」や「俺」といった頻出語によって生成結果に影響が生じたことから，生成性能向上のために，単に名詞を生成を生成に利用するのではなく話題を提供しやすい語を見つける必要がある．特定の語を抽出する方法として固有表現抽出 [11][12] という手法が存在しており，この手法では学習によって固有名詞や公共の組織，施設名のみを抽出することが可能とされている．固有表現抽出を用いることでトピックモデルを用いた場合に問題となっていた，曜日や時間に関する語を辞書から除去すると同時に固有表現を持つ文のみを抽出，テンプレート化することが可能となるだろう，また，テンプレート自体が適切でないという問題も考慮できると共に不適な文の生成抑制にも繋がる可能性が考えられる．今回の実験では発話タイプ推定から返答に利用するテンプレートが選択されたが，発話タイプの推定によってどの程度テンプレート選択の尤もらしさが向上したか評価を行われてはいなかった．テンプレート選択精度が生成文に対して適当な返答文生成のために必要な要素であると確認されたため，発話タイプの推定に関しても今後評価を行っていく必要があると考えられる．また返答文作成時，本研究では文法そのものに対する考慮を行っていなかったが，長文のテンプレートが選択される場合には，文書内の動詞との係り受け関係，主語もしくは目的語として適切な語が選択されているかを語の選択時点で精査し，適した語を選ぶことで不適な文の生成を抑制する可能性についても考える必要がある．

第5章　結論

本研究はテンプレートを用いた発話生成手法に対して，Support Vector Machine (SVM) を用いた発話タイプ推定に基づくテンプレート選択，N-gram モデルによる単語連接関係考慮，トピックモデルを加えることで，文構造を維持しつつ話題考慮が可能であると考え，従来手法と提案手法の比較実験・評価による検討を行った．実験アンケート結果より，連接関係の考慮が尤もらしい文の生成のためには必要であることが確認された．個別の生成結果では時間に関する語や話題に直接関係しない頻出語を削除することで，よりよい生成を行うことができる可能性が示唆された．テンプレートについても固有表現を持つ文を抽出することや，文構造そのものに注目し，主語や目的語として適切な語を選出するといったテンプレート作成手法，文生成手法の検討が必要だと考えられる．固有表現抽出に関しては学習データを作成する必要があることから，文構造そのものを学習データとすることで返答に用いられやすい文のみを抽出することができる可能性についても検証の必要がある，

今後の展望としては，より尤もらしい文生成のために考慮すべき連接数の考慮や発話タイプ推定による発話テンプレート性能の評価と雑談発話に対して適切なタグ種類の検討と評価，固有表現抽出を用いた場合のテンプレート自動生成及び辞書作成による生成文の尤もらしさ評価，文構造解析による適切な語の選択手法の検討を行っていく．

謝辞

研究を見守り，助言と支援を惜しまずに提供してくださった，若林先生，手塚先生に感謝を捧げると共に，入学から充実した学校生活を提供して下さった先生方，学友にも感謝いたします．ありがとうございます．

参考文献

[1] 小磯 花絵，石本 祐一，菊池 英明．”大規模日常会話コーパスの構築に向けた取り組み: 会話収録法を中心に”．言語・音声理解と対話処理研究会．2015，vol. 74，p. 37-42.

[2] 柴田 雅博，冨浦 洋一，西口 友美．”雑談自由対話を実現するための WWW 上の文章からの妥当な候補文選択手法”．人工知能学会論文誌．2009，vol. 24，no. 24，p. 507-519.

[3] 稲葉 通将，神園 彩香，高橋 健一．”Twitter を用いた非タスク指向型対話システムのための発話候補文獲得”．人工知能学会論文誌．2014，vol. 29，no. 1，p. 21-31.

[4] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, Yoshihiro Matsuo. *”Towards an open-domain conversational system fully based on natural language processing”*, In Proc. COLING．2014，p. 928-939.

[5] 安藤 秀哲，高橋 勇，黒岩 丈介，小高 知宏，小倉 久和．”チャットにおける会話の特徴と会話エージェントの検討”．福井大学工学部研究報告．2002，vol. 50，no. 2，p. 173-180.

[6] Suykens，Johan AK，Joos Vandewalle．*”Least squares support vector machine classifiers”*．Neural processing letters．1999，vol. 9，p. 293-300.

[7] Terrence S. Furey，Nello Cristianini，Nigel Duffy，David W. Bednarski，Michel Schummer David Haussler．*”Support vector machine classification and validation of cancer tissue samples using microarray expression data”*，Bioinformatics．2000，vol. 16，no. 10，p. 906-914.

[8] Jurafsky Dan，Elizabeth Shriberg，Debra Biasca．*”Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual”*．Institute of Cognitive Science Technical Report．1997，p. 97-101.

[9] David M. Blei，Andrew Y. Ng，Michael I. Jordan．*”Latent dirichlet allocation”*．Journal of Machine Learning Research．2003，vol. 3，p. 993-1022.

[10] Mimno. David，Matt Hoffman，David Blei．*”Sparse stochastic inference for latent Dirichlet allocation”*．arXiv．2012，https://arxiv.org/abs/1206.6425.

[11] 坪井 祐太，森 信介，鹿島 久嗣，小田 裕樹，松本 裕治．”日本語単語分割の分野適応のための部分的アノテーションを用いた条件付き確率場の学習”，情報処理学会論文誌．2009，vol. 50，no. 6，p. 1622-1635.

[12] 鈴木 雅之，黒岩 龍，印南 圭祐，小林 俊平，清水 信哉，峯松 信明，広瀬 啓吉．”条件付き確率場を用いた日本語東京方言のアクセント結合自動推定”．電子情報通信学会論文誌．2013，vol. 96，no. 3，p. 644-654.